chinaXiv:201805.00473v1

基于分位函数的直方图符号数据非负主成分分析法 *

李竹婷，陈秀宏，孙慧强

(江南大学 数字媒体学院, 江苏 无锡 214122)

摘　要：针对已有的符号数据主成分分析法大都采用部分代表性信息来代替符号数据的缺点，提出一种直方图符号数据的主成分分析法。直方图数据以概率分布的形式表示符号数据，更全面准确。根据直方图数据特点将其用分位函数表示，引入充分考虑直方图数据概率分布的 Wasserstein 距离，计算直方图变量协方差矩阵，从而进行主成分分析。但该方法求得的前若干个最大特征所对应的特征向量不一定为非负的，这样在用分位函数表示主成分时不能保证它也是分位函数。为此，又结合 Dias[1]等人的 DSD(distribution and symmetric distribution)回归模型，对每个直方图变量定义相应的对称分布变量，根据 Wasserstein 距离下的广义协方差矩阵得到具有非负系数的所有主成分。通过实验说明了该算法的有效性。该方法同时克服了文献[2]中直方图 PCA 系数可能为负的缺点，更多地保留了原始数据的信息。

关键词：主成分分析；直方图数据；分位函数；Wasserstein 距离；协方差矩阵

中图分类号：TP391.4　　doi: 10.3969/j.issn.1001-3695.2018.03.0151

Principal component analysis of histogram data with non-negative coefficients based on quantile function

Li Zhuting, Chen Xiuhong, Sun Huiqiang

(*School of Digital Media*, *Jiangnan University*, *Wuxi Jiangsu* 214000, *China*)

Abstract: Since the existing principal component analysis(PCA) of symbolic data mostly use some representative information instead of symbolic data, a histogram principal component analysis is proposed. Represent a histogram data by a quantile function with its characteristic, and introduce the Wasserstein distance which fully takes into account the probability distribution of the histogram data. It is easy to obtain the covariance matrix to perform the principal component analysis using this distance. However, the eigenvectors corresponding to the first m largest eigenvalues obtained by this method is not necessarily negative, so it cannot guarantee that the principal components are also quantile functions when they are represented by the quantile functions. For this point, combining the idea of DSD (distribution and symmetric distribution) regression model studied by Dias [1]et al, defining the corresponding symmetric distribution variables for each histogram variable, then obtain the non-negative principal component coefficients with the generalized covariance matrix. The experiments show the effectiveness of the algorithm. Besides, this method overcomes the disadvantage that the PCA coefficient of the histogram in [2] may be negative and retains more information of the original data.

Key words: Principal component analysis; histogram data; the quantile function; Wasserstein distance; covariance matrix

0　引言

随着“大数据”时代的到来，符号数据有着越来越广泛的应用，其中最具代表性的是区间符号数据和直方图符号数据。对于区间符号数据，最著名的有顶点主成分分析法（VPCA）和中点主成分分析法（CPCA），这两种方法均将一个区间型数据看作一个超立方体，分别用超立方体的顶点和中点来代表整个超立方体的信息。后来，Wang 等提出了全信息主成分分析法（CIPCA）[3]以及关于正态分布的主成分分析法（ND-PCA）[4]，这两种方法分别假设区间数据呈均匀分布和正态分布，计算区间型数据的协方差矩阵，以此进行主成分分析。然而，这两种方法都是基于假设区间型数据服从某种分布的，不具有普遍性。直方图数据可以看作是对区间型数据内部进行统计分析的结果，因此可以表示任意不规则分布的区间型数据。

由于直方图数据可以看做是一个分布，因此计算起来也比区间型数据复杂。在已有的直方图数据主成分分析法中，很多

收稿日期：2018-03-05；**修回日期：**2018-04-19　**基金项目：**国家自然科学基金资助项目（61373055）；2017 年江苏省研究生科研创新计划资助项目（KYCX17_1500）

作者简介：李竹婷（1993-），女，山西运城人，硕士研究生，主要研究方向为模式识别、数据挖掘（18800585276@163.com）；陈秀宏（1964-），男，教授，博士（后），主要研究方向为数字图像处理、模式识别；孙慧强（1993-），男，硕士研究生，主要研究方向为模式识别.

算法与区间型数据类似。Rodriguez 等人[5]将直方图数据转换为区间型数据来进行计算。Makosso-Kallyth 和 Diday[6]提出了一种定义直方图数据平均值的方法，用平均值来代替整个直方图数据，该方法同许多区间型主成分分析法类似，也是采用部分代表性信息来代替整个直方图数据变量的信息。此外，Nagabhushan 和 Kumar[2]定义了单位直方图矩阵并通过此类直方图矩阵的加减乘除运算来求得协方差矩阵并由此获得主成分，但是此方法求得重构后的直方图可能会出现负值，与实际情况不符，也因此会丢失大量的信息。

本文首先根据直方图数据的特点提出一种直方图数据的分位函数表示形式，这种表示形式大大减少了直方图数据计算的复杂度。然后在分位函数的基础上定义了 Wasserstein 距离，该距离充分利用直方图数据的概率分布进行计算，与其他只利用区间端点信息的距离相比，对直方图数据间的度量更准确。通过该距离可以求出一组直方图变量的中心直方图以及协方差矩阵，如直接利用该协方差矩阵进行主成分分析，此时的表示系数不一定全非负，而分位函数为非递减函数，所以分位函数线性表示不一定是分位函数。为此，借助 Dias[1]等人的思想，对每个直方图变量定义对称分布变量，对以上主成分进行修正。该方法不但解决了以往算法中只利用符号数据的部分信息来计算的缺点，保留了更多原始信息，更具有普遍性，同时克服了文献[2]中重构直方图权重可能为负这一缺陷。通过模拟数据以及 2010 年股票数据验证了本文算法的有效性。

1　直方图数据以及 Wasserstein 距离

1.1　直方图的定义及其相关算法

假设 Y 为一个直方图变量，如图 1 所示。其所在区间为 $S=[\underline{y},\overline{y}]$。将 S 划分为 H 个相继的区间 $\{I_1,I_2,\cdots,I_h\}$，其中 $I_h=[\underline{y_h},\overline{y_h}]$，$h=1,2,\cdots,H$，$I_h\cap I_k=\varnothing\ (h\neq k)$ 且 $\bigcup_{h=1}^{H}I_h=S$。于是得到 Y 的直方图表示：$Y=\{(I_1,f_1),(I_2,f_2),\cdots,(I_H,f_H)\}$。其中 $0\le f_i\le 1$，且 $\sum_{i=1}^{H}f_i=1$。定义 Y 的累计权为

$$w_l=\begin{cases}0, & l=0\\ \sum_{h=1}^{l}f_h, & l=1,2,\cdots H\end{cases}$$

则 Y 的经验分布函数为

$$F(y)=w_{l-1}+(y-\underline{y_l})\frac{w_l-w_{l-1}}{\overline{y_l}-\underline{y_l}},\qquad \underline{y_l}\le y\le\overline{y_l} \tag{1}$$

那么，它的逆函数即其分位函数（Quantile Function）为

$$Q(t)=F^{-1}(t)=\underline{y_l}+\frac{t-w_{l-1}}{w_l-w_{l-1}}(\overline{y_l}-\underline{y_l})\qquad w_l\le t\le w_{l-1} \tag{2}$$

由于分位函数是分段函数，因此，将直方图数据以分位函数的形式表示，降低了直方图数据计算的复杂度，更方便计算。但分段函数进行运算时，需要具有相同的分段数与分段区间，因此，在对直方图的分位函数进行计算时，需要将其重新构造使参加计算的分位函数具有相同的分段（其对应的直方图数据也被重新构造使得所有直方图被分成相同段数的子区间，并且对应的同一子区间上权值相等）。

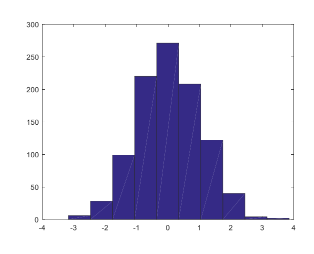


图 1　直方图数据

1.2　两个直方图间距离的表示

为计算两个直方图变量 Y_i 与 Y_j 间的距离，首先需要对其进行重构使其对应的分位函数具有相同的分段。根据 Irpino[17]的方法，将 Y_i 与 Y_j 的累积权 $w^i=\{w_0^i,w_1^i,\cdots,w_{n_i}^i\}$ 和 $w^j=\{w_0^j,w_1^j,\cdots,w_{n_j}^j\}$ 进行合并并按照从小到大的顺序进行排列得到集合 $w^{ij}=\{w_0^{ij},w_1^{ij},\cdots,w_{n_{ij}}^{ij}\}$，其中 $w_0=0$，$w_{n_{ij}}^{ij}=1$，且 $\max(n_i,n_j)\le n_{ij}\le(n_i+n_j-1)$，此时，对每组权 w_{l-1} 和 w_l 可确定两个区间：

$$I_{li}=\left[Q_i(w_{l-1}),Q_i(w_l)\right],\quad I_{lj}=\left[Q_j(w_{l-1}),Q_j(w_l)\right]。$$

因此，直方图 Y_i 与 Y_j 的分位函数 Q_i 与 Q_j 被重写为具有相同分段 $w^{ij}=\{w_0^{ij},w_1^{ij},\cdots,w_{n_{ij}}^{ij}\}$ 的分位函数，直方图 Y_i 与 Y_j 也被表示为每个子区间具有相同权重的直方图。此时，便可对两个具有相同分段的分位函数进行计算。

定义两个直方图变量 Y_i 与 Y_j 间的 Wasserstein 距离为

$$d_w^2(Y_i,Y_j)=\int_0^1\left[Q_i(t)-Q_j(t)\right]^2dt=\sum_{l=1}^{n_{ij}}\int_{w_{l-1}}^{w_l}\left[Q_i(t)-Q_j(t)\right]^2dt \tag{3}$$

其中：每组权 w_{l-1} 和 w_l 对应 Y_i 与 Y_j 的两个区间的中心和半径分别为

$$c_{lu}=\frac{Q_u(w_{l-1})+Q_u(w_l)}{2},\quad r_{lu}=\frac{Q_u(w_l)-Q_u(w_{l-1})}{2},\quad u=i,j$$

于是得到以下结果：

命题 1

$$d_w^2(Y_i,Y_j)=\sum_{l=1}^{n_{ij}}f_l\left[(c_{li}-c_{lj})^2+\frac{1}{3}(\mathrm{r}_{li}-r_{lj})^2\right] \tag{4}$$

$$\int_0^1 Q_i(t)Q_j(t)dt=\sum_{l=1}^{n_{ij}}f_l\left[c_{li}c_{lj}+\frac{1}{3}r_{li}r_{lj}\right] \tag{5}$$

其中 $f_l=w_l-w_{l-1}$，$w_l,w_l\in W^{ij},l=1,2,\cdots,n_{ij}$。

1.3　中心直方图的求解

给定的 n 个直方图 $Y_1,Y_2,\cdots,Y_n$，其中心 Y_b 也是一直方图变量。将 $Y_1,Y_2,\cdots,Y_n$ 的累积权进行合并并按从小到大的顺序排列成一集合 W，并记该集合中元素个数为 m，则由命题 1 可知求中心直方图即为极小化以下函数：

$$\begin{aligned}f\left(c_{1b},r_{1b},\cdots,c_{mb},r_{mb}\right)&=\sum_{i=1}^{n}d_w^2\left(Y_i,Y_b\right)\\&=\sum_{i=1}^{n}\sum_{l=1}^{m}f_l[(c_{li}-c_{lb})^2+\frac{1}{3}(r_{li}-r_{lb})^2]\end{aligned} \tag{6}$$

其中：c_{lb} 与 r_{lb} 为中心直方图基本区间的中心与半径。解得：

chinaXiv:201805.00473v1

ChinaXiv合作期刊

$$c_{lb}=\frac{1}{n}\sum_{i=1}^{n}c_{li}\ ,\quad r_{lb}=\frac{1}{n}\sum_{i=1}^{n}r_{li}\ ,\quad l=1,2,\cdots,m\ 。\tag{7}$$

从而 n 个直方图数据的中心直方图表示为

$$\begin{aligned}Y_b=\{&([c_{1b}-r_{1b},c_{1b}+r_{1b}],f_1),\cdots,\\&([c_{lb}-r_{lb},c_{lb}+r_{lb}],f_l),\cdots,([c_{mb}-r_{mb},c_{mb}+r_{mb}],f_m)\}\end{aligned}\tag{8}$$

1.4 方差与协方差

设 $X_1,X_2,\cdots,X_p$ 为 p 个相互独立的直方图变量，其中 X_j 中的 n 个元素 $X_{1j},X_{2j},\cdots,X_{pj}$ 均为直方图变量且相互独立。

定义两个直方图符号变量 X_k 和 X_j 的标量积为

$$X_k^TX_j=\sum_{j=1}^{n}\int_0^1 Q_{ik}(t)Q_{ij}(t)dt\tag{9}$$

其中：$Q_{ik}(t)$ 与 $Q_{ij}(t)$ 分别为直方图变量 $X_{ik}(t)$ 与 $X_{ij}(t)$ 的分位函数。

记 $\overline{Q_j(t)}$ 为直方图变量 X_j 的中心直方图的分位函数表示形式，根据 Wasserstein 距离，X_j 的方差为

$$Var(X_j)=\frac{1}{n}\sum_{i=1}^{n}d_w^2(X_{ij},\overline{X_j})=\frac{1}{n}\sum_{i=1}^{n}\int_0^1[Q_{ij}(t)-\overline{Q_j(t)}]^2dt\tag{10}$$

标准差为 $STD(X_j)=\sqrt{Var(X_j)}$，

从而 X_{ij} 的标准化偏差为

$$SD_{ij}(t)=\frac{Q_{ij}(t)-\overline{Q_j(t)}}{STD(X_j)}=\frac{Q_{ij}(t)-\overline{Q_j(t)}}{\sqrt{Var(X_j)}}\ ,\quad 0\le t\le 1\tag{11}$$

可以证明以下结果成立：

命题 2　$\sum_{i=1}^{n}[Q_{ij}(t)-\overline{Q_j(t)}]=0$ ，　$\frac{1}{n}\sum_{i=1}^{n}\int_0^1[SD_{ij}(t)]^2dt=1$ 。

X_j 与 X_k 的协方差为

$$COVAR(X_j,X_k)=\frac{1}{n}\sum_{i=1}^{n}\int_0^1(Q_{ij}(t)-\overline{Q_j(t)})(Q_{ik}(t)-\overline{Q_k(t)})dt\tag{12}$$

则由命题 2 得

$$COVAR(X_j,X_k)=\frac{1}{n}\sum_{i=1}^{n}\int_0^1 Q_{ij}(t)Q_{ik}(t)dt-\int_0^1\overline{Q_j}(t)\overline{Q_k}(t)dt\tag{13}$$

2 直方图数据非负主成分分析法

2.1 经典主成分分析法向直方图数据的推广

根据经典主成分分析法，直方图数据 $X_1,X_2,\cdots,X_p$ 的主成分 Y 定义为它们的线性组合，即 $Y=Xu=\sum_{j=1}^{p}u_jX_j$ ，其中 $u=(u_1,u_2,...u_p)^T$ 满足 $u^Tu=1$ 。那么主成分 Y 的方差为：

$VAR(Y)=VAR(\sum_{j=1}^{p}u_jX_j)=u^TDu$ ，其中 D 为 $X_1,X_2,\cdots,X_p$ 的协方差矩阵

$$D_{ij}=\begin{cases}VAR(X_i)\ , & i=j\\ COVAR(X_i,X_j), & i\ne j\end{cases},$$

所以求主成分转换为条件 $u^Tu=1$ 下极大化方差 $VAR(Y)$ 的优化问题：

$$\begin{aligned}&\max\quad u^TDu\\&s.t.\quad u^Tu=1\end{aligned}\tag{14}$$

此问题的求解可以转换为求协方差矩阵 D 的特征值和特征向量。记 λ_k 为 D 的第 k 个最大特征值，对应的特征向量为 u_k ，其中 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_r$ ，$r=rank(D)$ ，则直方图数据样本矩阵的第 k 个主成分表示为

$$Y_k=Xu_k=\sum_{j=1}^{p}u_{kj}X_j\ ,\quad k=1,2,\cdots,r\ 。$$

于是，得到以下直方图数据的主成分分析算法：

a)根据式(10)和(13)计算直方图数据 $X_1,X_2,\cdots,X_p$ 的协方差矩阵 D；

b)求解特征方程 $Du=\lambda u$ 的前 m 个最大特征值 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_m$ 及对应的正交特征向量 $u_1\ge u_2\ge\cdots\ge u_m$，$m\le p$；

c)计算第 k 个主成分 $Y_k=Xu_k=\sum_{j=1}^{p}u_{kj}X_j$ ，$k=1,2,\cdots,m$，$m\le r$ 。

2.2 基于分位数的非负直方图主成分分析法

在上述过程中，对协方差矩阵进行求解，所得到的特征值与特征向量不能保证全部非负，因此在对直方图进行重构计算主成分直方图时可能会出现问题。另外，主成分 Y_k 也可以用分位函数表示：

$$Q_{ik}^Y(t)=\sum_{j=1}^{p}u_{jk}Q_{ij}(t)\ ,\quad i=1,2,\cdots,n\ ,\quad 0\le t\le 1.\tag{15}$$

由于分位函数是单调增加的，且只有单调增加函数的正线性组合才是单调增加的，所以为保证主成分 $Q_{ik}^Y(t)$ 也是分位函数，参数 u_{jk} 也必须是非负的，即 $u_{jk}\ge 0$，$j=1,2,\cdots,p$ 。

Dias 等人[1]给出了一种非负约束下基于分位函数间 Wasserstein 距离的回归方法，通过在回归表达式中增加对称分布的分位函数而扩充了回归因子的个数。以下利用该思想研究非负约束的主成分分析法并给出主成分分位函数的修正形式。

假设随机变量 X_{ij} 的经验或理论概率密度函数为 f_{ij}（其分位函数为 Q_{ij}），其对应的对称分布 $\tilde{f}_{ij}$（分位函数为 $\tilde{Q}_{ij}$）是将 f_{ij} 的支撑乘以-1 并使得两个分位函数和的积分为零，即 $\int_0^1\left[Q_{ij}(t)+\tilde{Q}_{ij}(t)\right]dt=0$ 。于是主成分 Y 的分位函数表示的修正形式为

$$Q_{ik}^Y(t)=\sum_{j=1}^{p}u_{jk}Q_{ij}(t)+\sum_{j=1}^{p}\tilde{u}_{jk}\tilde{Q}_{ij}(t)\ ,\tag{16}$$

其中: $u_{jk}\ge 0$，$\tilde{u}_{jk}\ge 0$ 。

记与 X_{ij} 对应的对称分布的变量为 $\tilde{X}_{ij}$，其对应的的分位函数表示为 $\tilde{Q}_{ij}$，变量 $\tilde{X}_j=[\tilde{X}_{1j},\tilde{X}_{2j},\cdots,\tilde{X}_{nj}]^T$ ($j=1,2,\cdots,p$)为 n 维随机向量，那么原始变量组 $X=[X_1,X_2,\cdots,X_p]$ 对应的对称分布为 $\tilde{X}=[\tilde{X}_1,\tilde{X}_2,\cdots,\tilde{X}_p]$。假设向量组 $\overline{X}=[X_1,X_2,\cdots,X_p,\tilde{X}_1,\tilde{X}_2,\cdots,\tilde{X}_p]=[X,\tilde{X}]$ 的协方差矩阵为 V，于是在非负约束条件（即 $v\ge 0$）下极大化方差 v^TVv 可表示为以下优化问题：

$$\begin{aligned}&\max\quad v^TVv\\&s.t.\quad v^Tv=1,\quad v\ge 0\end{aligned}\ ,\tag{17}$$

其中 $v=[v_1,v_2,\cdots,v_p,v_{p+1},v_{p+2},\cdots,v_{2p}]^T$ 。由于该问题具有非负约束，所以不能直接将之转换为求协方差矩阵 V 的特征值与特征向量。但它本身是一个二次优化问题，故可通过非线性优化算法来求解。假设 $v^{(1)}$ 是其最优解，则第一个主成分为

$Y^{(1)}=Xv^{(1)}=\sum_{j=1}^{p}v_{1j}X_j+\sum_{j=1}^{p}\tilde{v}_{1j}\tilde{X}_j$ ，其残差为

$$\overline{X}^{(1)}=\overline{X}^{(0)}-Y^{(1)}v^{(1)T}\ ,\tag{18}$$

chinaXiv:201805.00473v1

ChinaXiv合作期刊
chinaXiv:201805.00473v1

其中：$\overline{X}^{(0)}=\overline{X}$，计算$\overline{X}^{(1)}$的协方差矩阵$V_1$，用$V_1$替换问题（17）中的 V 并求得 $v^{(2)}$，从而有第二个主成分 $Y^{(2)}=\overline{X}^{(1)}v^{(2)}$。重复以上过程，得以下迭代式：

$$\overline{X}^{(k)}=\overline{X}^{(k-1)}-Y^{(k)}v^{(k)^T}, \tag{19}$$

再计算$\overline{X}^{(k)}$的协方差矩阵V_k并替换问题(17)中的 V 并解得 $v^{(k+1)}$，从而第 k 个主成分为$Y^{(k+1)}=\overline{X}^{(k)}v^{(k+1)}$。如此下去，即可获得$\overline{X}$的所有 r 个主成分，其中$r=rank(V)$。

以上方法最重要的特性是所有$v^{(1)},v^{(2)},\cdots,v^{(r)}$均为非负的且线性无关，但它们不必正交，且在 Wasserstein 距离下$\overline{X}$的最佳逼近为

$$\hat{\overline{X}}=\sum_{k=1}^{r}Y^{(k)}v^{(k)T} \tag{20}$$

如何直接由$\overline{X}$来确定主成分才能使$\hat{\overline{X}}$成为最佳逼近？考虑以下最小误差优化问题：

$$\begin{aligned}&\min\quad \varepsilon=\sum_{i=1}^{n}\sum_{j=1}^{2p}d_W^2(\overline{X}_{ij},\hat{\overline{X}}_{ij})\\ &s.t.\quad u_{ik}\ge 0,\tilde{u}_{ik}\ge 0,i=1,2,\cdots,n\ k=1,2,\cdots,r,\end{aligned} \tag{21}$$

其中：

$$d_W^2(\overline{X}_{ij},\hat{\overline{X}}_{ij})=\int_0^1\left\{Q_{ij}(t)-\sum_{k=1}^{r}v_{jk}\left[\sum_{s=1}^{p}u_{sk}Q_{ij}(t)+\sum_{s=1}^{p}\tilde{u}_{sk}\tilde{Q}_{is}(t)\right]\right\}^2dt$$
$$+\int_0^1\left\{\tilde{Q}_{ij}(t)-\sum_{k=1}^{r}v_{p+j,k}\left[\sum_{s=1}^{p}u_{sk}Q_{is}(t)+\sum_{s=1}^{p}\tilde{u}_{sk}\tilde{Q}_{is}(t)\right]\right\}^2dt$$
$$i=1,2,\cdots,n,\quad j=1,2,\cdots,p$$

$\overline{X}_{ij}$ 和 $\hat{\overline{X}}_{ij}$ 分别为 $\overline{X}$ 和 $\hat{\overline{X}}$ 中的第 (i,j) 个元素。如果 $(u_{1k},u_{2k},\cdots,u_{pk},\tilde{u}_{1k},\tilde{u}_{2k},\cdots,\tilde{u}_{pk})$为(21)的最优解，那么第 k 个主成分 $Y^{(k)}$ 的第 i 个元素的分位函数 $Q_{ik}{}^{Y}(t)$ 为

$$Q_{ik}^{Y}(t)=\sum_{j=1}^{p}u_{jk}Q_{ij}(t)+\sum_{j=1}^{p}\tilde{u}_{jk}\tilde{Q}_{ij}(t),$$
$$k=1,2,\cdots,r,\quad i=1,2,\cdots,n。$$

下面给出改进的直方图数据主成分分析法，称为非负直方图 PCA（NHV-PCA）。

a)根据式(10)和(13)计算出协方差矩阵 V。

b)利用式(17)和(19)计算非负向量组$v^{(1)},v^{(2)},\cdots,v^{(k)}$，其中 k 满足前 k 个主成分的累积方差贡献率达到一确定的百分比。

c)求解问题式(21)得最优解 $\mathbf{u}^{(k)}=(u_{1k},u_{2k},\cdots,u_{pk},\tilde{u}_{1k},\tilde{u}_{2k},\cdots,\tilde{u}_{pk})$，并计算第 l 个主成分中第 i 元素的分位函数：

$$Q_{il}^{Y}(t)=\sum_{j=1}^{p}u_{jl}Q_{ij}(t)+\sum_{j=1}^{p}\tilde{u}_{jl}\tilde{Q}_{ij}(t),$$
$$l=1,2,\cdots,k,\quad k\le r,\quad i=1,2,\cdots,n。$$

于是，计算$Q_{il}^{Y}(t)$的积分（称之为主成分$Q_{il}^{Y}(t)$的分位值）

$$q_{il}^{Y}=\int_0^1Q_{il}^{Y}(t)dt=\sum_{j=1}^{p}(u_{jl}-\tilde{u}_{jl})\int_0^1Q_{ij}(t)dt,\quad l=1,2,\cdots,k,$$
$$k\le r,\ i=1,2,\cdots,n。$$

3　数值实验结果与分析

3.1　模拟数据集

本小节利用模拟数据集验证算法的有效性。结合 Monte Carlo 实验方法，从每个符号数据内部随机选取 m 个单值数据来近似模拟该符号数据，m 越大，所取的这些数越能代表该符号数据。参考文献[3]，假设有三组不同类型的正态分布 C_1、C_2 和 C_3，它们的均值μ_j与方差δ_j服从不同区间上的均匀分布，如表 1 所示。

表 1　三类数据

	μ_j	δ_j
C1	U[-5,-2]	U[1,2]
C2	U[-3,3]	U[1,2]
C3	U[2,5]	U[1,2]

记$X_{n\times p}$为符号型数据样本矩阵，其中元素X_{ij}服从正态分布$N(\mu_{ij},\sigma_{ij})$。为了将本文算法与经典 PCA 和 ND-PCA 进行比较，需要生成 ND-PCA 法所需的正态分布值采样矩阵$X_{n\times p}$、经典 PCA 的单值采样矩阵以及适用于本文算法的直方图数据矩阵。数据矩阵生成的过程如下：

a）生成三个服从正态分布的符号数据样本矩阵$X_{n\times p}$，其第 j 个列向量$X_j=(X_{1j},X_{2j},...,X_{nj})^T$为第 j 个变量，含有 n 个观察值。对第 j 个正态分布变量，生成一个均值向量$\mu_j=(\mu_{1j},\mu_{2j},...,\mu_{nj})^T$和标准偏差向量$\sigma_j=(\sigma_{1j},\sigma_{2j},...,\sigma_{nj})^T$，这里$\mu_{ij}$和$\delta_{ij}$分别服从于$[a,b]$和[c,d]上的均匀分布（a,b,c,d 是任意的）。将这三个矩阵合并为一个 $3n\times p$ 矩阵$X_{3n\times p}$，执行 ND-PCA 算法并获得分类精度。

b）对符号数据矩阵$X_{3n\times p}$，从每个服从正态分布$N(\mu_{ij},\sigma_{ij}^2)$的元素中任意抽取 M 个数据，形成一个$(3n*M)\times p$的单值数据矩阵$K_{(3n*M)\times p}$，执行经典 PCA 并计算分类精度。

c）对符号数据矩阵$X_{3n\times p}$，从每个服从正态分布$N(\mu_{ij},\sigma_{ij}^2)$的元素中任意抽取 M 个数据并进行统计生成直方图数据，得到直方图数据矩阵$H_{3n\times p}$并执行 NHV-PCA 算法并分类精度。

以上三个实验均重复 R 次，并求出 R 次的平均值。所有实验均假设 n=50，p=6，R=10，而 M 则分别取 100，500，1000，5000，10000。实验结果如表 2 所示。

表 2　分类精度对比　　/%

M	对比方法	特征 1	特征 2	特征 3	特征 4	特征 5
	ND-PCA	84.13	82.93	90.4	93.86	99.07
100	经典 PCA	99.52	99.63	99.26	99.64	98.71
	NHV-PCA	98.48	98.86	98.67	98.86	98.5
500	ND-PCA	84.13	82.93	90.4	93.86	99.07

ChinaXiv合作期刊

	经典 PCA	99.32	98.95	98.05	99.6	98.93
	NHV-PCA	98.27	98.3	98	98.4	98.43
	ND-PCA	84.13	82.93	90.4	93.86	99.07
1000	经典 PCA	99.04	99.4	97.84	98.61	98.69
	NHV-PCA	98.67	98.86	96.57	97.71	98.1
	ND-PCA	84.13	82.93	90.4	93.86	99.07
5000	经典 PCA	98.64	98.93	97.62	98.95	98.99
	NHV-PCA	98.67	98.85	97.43	98	98.56
	ND-PCA	84.13	82.93	90.4	93.86	99.07
10000	经典 PCA	98.62	98.7	98.01	98.7	98.75
	NHV-PCA	98.66	98.69	97.93	98.69	98.73

表 3　时间对比　　/s

M		特征 1	特征 2	特征 3	特征 4	特征 5
	ND-PCA	0.84	0.83	0.90	0.94	0.99
100	经典 PCA	49.72	190.24	191.5	193.11	194.42
	NHV-PCA	244.82	243.5	245.69	250.79	264.1
	ND-PCA	0.84	0.83	0.90	0.94	0.99
500	经典 PCA	1312.85	5227.8	6248.99	5783.79	5561.13
	NHV-PCA	262.69	273.17	343.06	301.37	311.04
	ND-PCA	0.84	0.83	0.90	0.94	0.99
1000	经典 PCA	6757.77	7225	7247.12	7335.38	7658.65
	NHV-PCA	313.51	312.61	317.86	322.24	330.43
	ND-PCA	0.84	0.83	0.90	0.94	0.99
5000	经典 PCA	233509.3	234500.1	234525.1	234595.8	234607.1
	NHV-PCA	249.44	255.67	257.32	259.11	260.55
	ND-PCA	0.84	0.83	0.90	0.94	0.99
10000	经典 PCA	1284300	1284307	1284312	1284315	1284319
	NHV-PCA	289.32	297.57	310.32	314.66	317.49

由表 2 可见，随着 M 的逐渐增大，经典主成分分析法计算得到的分类精度逐渐降低。这是因为，当 M 较小时，所取得的样本只代表了符号数据内的部分信息，并且所取样本在符号数据内部比较分散，因此容易区分；当 M 较大时，每个符号数据内所采样的样本个数增加，它们在符号数据内部分散比较密集，且三类数据集本身有部分重叠元素，因此样本数据交叉重合的部分较多。另外，NHV-PCA 算法的分类精度相对稳定，这是因为，无论 M 取值为 100, 1000 或是 10000，NHV-PCA 算法都是对采样值进行统计形成直方图数据。另外，由表 2 还可以看到，随着 M 逐渐增大，经典 PCA 和 NHV-PCA 的分类精度趋于一致，从一定程度上说明 NHV-PCA 算法更能从整体上把握各类数据间的相关性。与前两种算法相比，ND-PCA 的分类精度在所取特征维数较低时也比较低，而在所取特征维数较高时会取得比较好的分类效果。但是，ND-PCA 方法仅适用于服从正态分布的符号数据，而本文提出的 NHV-PCA 算法则适用于具有任意分布的符号型数据，因此更具有普遍性。

表 3 给出了三种算法运行一次所消耗的时间。从表中结果可以看出，ND-PCA 所用时间最小，这是因为该算法只是利用了每个正态分布数据的均值 μ 和偏差 δ，并对协方差矩阵进行直接分解获得的特征向量，并且无论 M 取值为多少，ND-PCA 永远为 $3n\times p$ 的矩阵，计算量不变。而经典 PCA 虽然也是对协方差矩阵进行分解，但是当 M 较大时，所得到的数据越来越庞大，因此计算量也会随之增大。当 M 较小（例如 M=100）时，经典 PCA 运行一次所消耗的时间小于 NHV-PCA 法，而当数据较大（例如 M>=1000）时，经典 PCA 法的运行时间将远大于 NHV-PCA 法。例如，当 M=10000 时，经典 PCA 运行一次所用时间长达 1284300s（约为 14 天），而 NHV-PCA 法运行一次所用的时间大约为 290s 左右。虽然 NHV-PCA 法和经典 PCA 法运行一次所消耗的时间随着 M 的增大而增加，但整体上看 NHV-PCA 法的增大幅度远小于经典 PCA 的增大幅度。

由以上三种算法分类精度与时间的对比可以看到，当数据量非常庞大时，采用直方图主成分分析法可以快速而有效地提取出数据中的有用信息，并且能从全局上把握整个数据集。虽然 ND-PCA 花费时间也较短，但需要提取多个特征才能达到较

chinaXiv:201805.00473v1

高的分类精度，并且直方图算法相较于 ND-PCA 算法更具有普遍性，而不仅仅局限于正态分布型数据。

3.2　Iris 数据集

本小节利用 Iris 数据集与文献[2]中 Histogram PCA 算法进行对比。Iris 数据集由 150 个具有 4 个特征的样本组成。共有 3 个类，即 setosa，versicolour 和 verginica。其中，前 50 个样本属于第一类，中间 50 个样本属于第二类，最后 50 个样本属于第三类。从每个类别中随机抽取 30 个样本，对其归一化处理，再对其统计分析生成直方图数据，每个类别随机抽取 5 次，因此可以获得 15 个 4 维直方图数据。绘制直方图数据如图 2 所示。

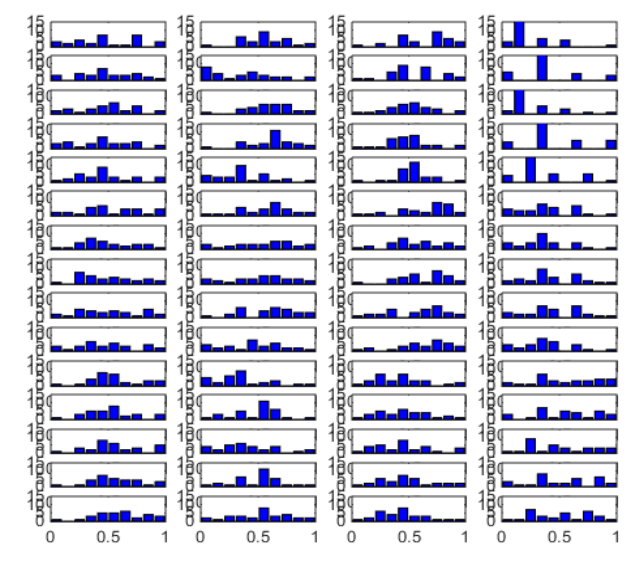


图 2　Iris 原始数据直方图

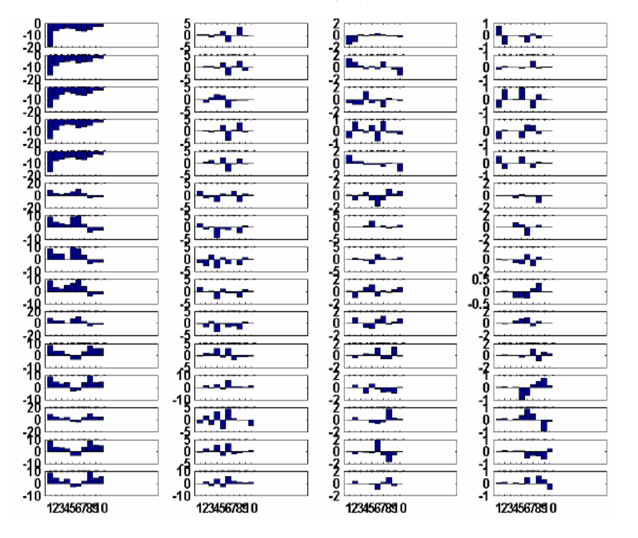


图 3　Histogram PCA 算法主成分直方图

图 3、4 分别绘制出经过 Histogram PCA 算法与非负直方图主成分分析法计算后得到的直方图。经过降维后的直方图数据可以很明显的根据前两维变量将三类数据分辨出来，都起到了很好的降维效果。但是对比图 3、4 发现，经过 Histogram PCA 算法得到的直方图数据许多概率为负，已经不属于传统意义上的直方图数据，与实际情况不符，但是经过非负直方图主成分分析法降维得到的直方图数据可以很完整的表示出一个直方图数据，具有很好的实际意义。

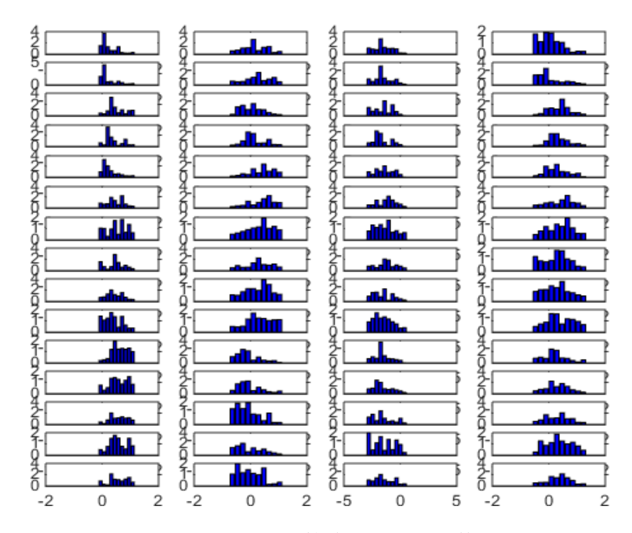


图 4　NHV-PCA 算法后的直方图数据

3.3　股票数据集

中国股票市场数据量庞大，结构复杂，由于股票的合并，重组，重新命名，以及暂停恢复等，导致对单支股票进行追踪研究较为困难，但如果从宏观上以每一类股票为研究对象，可以从整体上把握各类股票间的关系和规律。为此，本实验选取上海证券交易市场 2010 年 1 月 1 日到 2010 年 12 月 31 日所有上市公司的交易数据，选取 5 个变量：年个股总市值(X_1)、P/E 值(X_2)、换手率(X_3)，波动率(X_4)以及回报率(X_5)，将每类股票数据进行打包生成直方图数据进行实验。数据处理过程如下：

首先将所有股票按市值进行排序，并分为大盘股、中盘股和小盘股；其次，对每个类别股票按照市盈率(P/E）进行排序，去除掉每个类别中最高和最低的 5%后再取 P/E 的中位数作为临界点，P/E 大的部分为增长股，小的部分为价值股，因此形成六种股票：大盘增长股（L-G），大盘价值股（L-V），中盘增长股（M-G），中盘价值股（M-V）；小盘增长股（S-G），小盘价值股（S-V）。对六种股票的每个变量进行归一化处理并统计形成直方图变量，如图 5 所示，图中每一行代表一种股票，每一列代表一个变量。

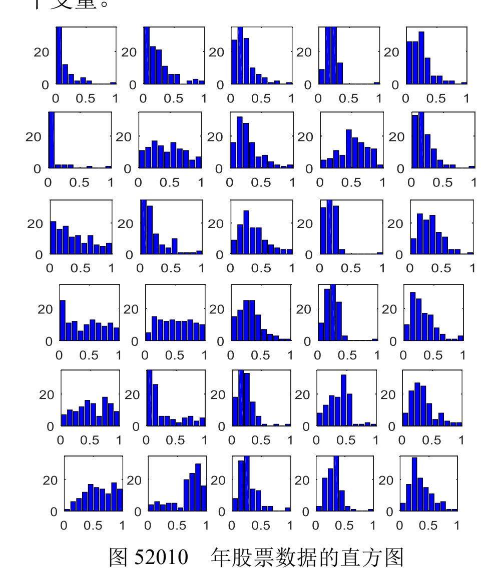


图 52010　年股票数据的直方图

利用本文提出的非负主成分分析 NHV-PCA 算法，在提取原数据中 61%的信息情况下可以得到前两个主成分：

chinaXiv:201805.00473v1

ChinaXiv合作期刊

$$Q_1^Y = 0.7381\tilde{Q}_1 + 0.1280\tilde{Q}_2 + 0.0326Q_3 + 0.3538Q_4$$

$$Q_2^Y = 0.5378Q_1 + 0.4753\tilde{Q}_2 + 0.6304\tilde{Q}_4 + 0.2312Q_5$$

图 6 绘制了前两个主成分的大致分布。观察发现，第一个主成分中 Q_1、Q_2 与 Q_3、Q_4 呈反比。其中，Q_1 代表一个公司的总市值，Q_2 对应其 P/E 值，用来衡量一支股票是否被高估；Q_3 和 Q_4 分别表示换手率和波动率，衡量一支股票的交换频率与价格波动，这两个变量代表了一支股票的动态变化因此可以表示一支股票的“风险”。Q_1 与 Q_3、Q_4 的负相关表明中国股票市场大盘股风险低、小盘股风险高的现象。

另外由图 6 发现，第一个主成分中市值所占比重最大，第二个主成分中 P/E 值与波动率所占比重较大，因此第二个主成分可以近似表示“风险”。对 6 种类型的股票进行主成分重构后得到图 7 所示主成分直方图。

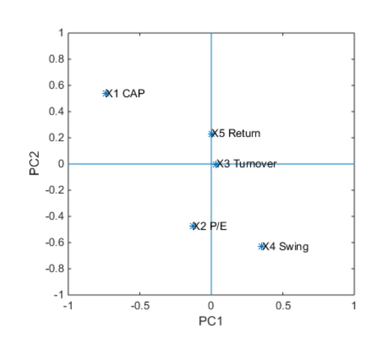


图 6　前两个主成份的大致分布

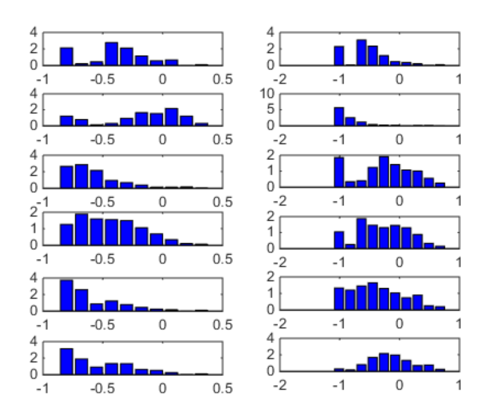


图 7　6 种类型的股票数据在第一与第二主成分上的分布

图 7 中第一列与第二列分别代表第一个主成分与第二个主成分 6 种股票的分布。每一行代表一种股票类型。由上至下分别为 L-G，L-V，M-G，M-V，S-G 和 S-V。由于第一个主成分主要代表了股票的市值，观察第一列直方图数据可发现其分布与实际吻合。第二个主成分代表风险，第二列直方图从上至下分布中心由小变大，验证了大盘股风险小，小盘股风险大的说法。

图 8 绘制出了对所有单支股票利用经典主成分分析法进行降维后提取前两维得到的结果。可以发现，6 种类型的股票杂乱的交叉在一起，无法从宏观上提取出六种股票之间的有效信息。因此可以说明，符号数据分析法可以从整体上把握研究对象关系，挖掘深层次规律的方法，具有重要的研究意义。

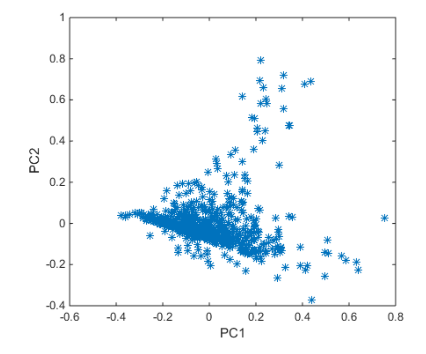


图 8　对所有单支股票利用经典 PCA 进行降维后的效果

4　结束语

在实际问题中符号型数据有着广泛的应用，从而出现了许多区间符号数据的降维方法。最为经典的方法有 C-PCA 和 V-PCA、CIPCA 和正态分布 ND-PCA 等，这些方法均假设区间型数据服从均匀分布或者正态分布，但是对于非均匀分布或非正态分布的数据具有一定的局限性。本文针对直方图符号数据利用 Wasserstein 距离和 Dias[8]等人的线性回归方法，给出了一种非负直方图主成分分析法，该方法相比已有的符号数据主成分分析算法更具有普遍性，并且克服了文献[3]中直方图 PCA 算法所获得主成分系数可能为负的缺点，更好地保留了此类数据的原始信息。通过模拟数据集和在中国股票市场的实证分析验证了算法的有效性。

参考文献：

[1] Dias S, Brito P. Linear regression model with histogram-valued variables [J]. Statistical Analysis & Data Mining the Asa Data Science Journal, 2015, 8 (2): 75–113.

[2] Nagabhushan P, Kumar R P. Histogram PCA [C]// Proc of International Symposium on Neural Networks. [S. l.] : Springer-Verlag, 2007: 1012-1021.

[3] Wang Huiwen, Guan Rong, Wu Junjie. CIPCA: Complete-information-based principal component analysis for interval-valued data [J]. Neurocomputing, 2012, 86 (4): 158-169.

[4] Wang Huiwen, Chen Meiling, Shi Xiaojun, *et al.* Principal component analysis for normal-distribution-valued symbolic data [J]. IEEE Trans on Cybernetics, 2016, 46 (2): 356-365.

[5] Brito P, Dias S. A new linear regression model for histogram-valued variables [C]// Proc of ISI World Statistics Congress. 2011.

[6] Verde R, Irpino A. Comparing histogram data using a Mahalanobis–Wasserstein distance [C]// Computational Statistics. Physica-Verlag, 2008: 77-89.

[7] Verde R, Irpino A, Balzanella A. Dimension reduction techniques for distributional symbolic data [J]. IEEE Trans on Cybernetics, 2016, 46 (2): 344.

[8] Irpino A, Verde R. Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein distance [J]. Advances in Data Analysis and Classification, 2015, 9 (1): 81-106.

chinaXiv:201805.00473v1

ChinaXiv合作期刊

[9] Long Wen, Mok H M Kenry, Hu Yan, *et al.* The style and innate structure of the stock markets in China [J]. Pacific-Basin Finance Journal, 2009, 17 (2): 224-242.

[10] Pachner A R, Delaney E, Ricalton N S. Generalization of the principal components analysis to histogram data [J]. Principles & Practice of Knowledge Discovery in Databases, 2000, 4 (6): 345-351.

[11] Verde R, Irpino A. Ordinary least squares for histogram data based on wasserstein distance [C]// Proc of COMPSTAT. Physica-Verlag, 2010.

[12] Ichino M. The quantile method for symbolic principal component analysis [J]. Statistical Analysis & Data Mining the Asa Data Science Journal, 2011, 4 (2): 184–198.

[13] 李汶华. 区间型符号数据分析理论方法及其在金融中的应用研究 [D]. 天津：天津大学, 2006. (Li Wenhua, Study of the theory & methodology of interval-valued symbolic data analysis with application to finance [D]. Tianjin: Tianjin University, 2006.)

[14] Sun M K, Diday E. Adaptation of interval PCA to symbolic histogram variables [J]. Advances in Data Analysis & Classification, 2012, 6 (2): 147-159.

[15] Dias S, Brito P. Linear regression model with histogram-valued variables [J]. Statistical Analysis & Data Mining the Asa Data Science Journal, 2015, 8 (2): 75-113.

[16] Sun M K. Principal axes analysis of symbolic histogram variables [J]. Statistical Analysis & Data Mining the Asa Data Science Journal, 2015, 9 (3): 188-200.

[17] Irpino A, Verde R. A new wasserstein based distance for the hierarchical clustering of histogram symbolic data [J]. Studies in Classification Data Analysis & Knowledge Organization. 2006: 185-192.

chinaXiv:201805.00473v1